SpikeGPT: Generative Pre-trained Language Model with Spiking Neural Networks

Year: 2023 | Citations: 111 | Authors: Rui Zhu, Qihang Zhao, J. Eshraghian

---

Abstract

As the size of large language models continue to scale, so does the computational resources required to run it. Spiking Neural Networks (SNNs) have emerged as an energy-efficient approach to deep learning that leverage sparse and event-driven activations to reduce the computational overhead associated with model inference. While they have become competitive with non-spiking models on many computer vision tasks, SNNs have also proven to be more challenging to train. As a result, their performance lags behind modern deep learning, and we are yet to see the effectiveness of SNNs in language generation. In this paper, inspired by the Receptance Weighted Key Value (RWKV) language model, we successfully implement `SpikeGPT', a generative language model with binary, event-driven spiking activation units. We train the proposed model on two model variants: 45M and 216M parameters. To the best of our knowledge, SpikeGPT is the largest backpropagation-trained SNN model to date, rendering it suitable for both the generation and comprehension of natural language. We achieve this by modifying the transformer block to replace multi-head self attention to reduce quadratic computational complexity O(N^2) to linear complexity O(N) with increasing sequence length. Input tokens are instead streamed in sequentially to our attention mechanism (as with typical SNNs). Our preliminary experiments show that SpikeGPT remains competitive with non-spiking models on tested benchmarks, while maintaining 20x fewer operations when processed on neuromorphic hardware that can leverage sparse, event-driven activations. Our code implementation is available at https://github.com/ridgerchu/SpikeGPT.